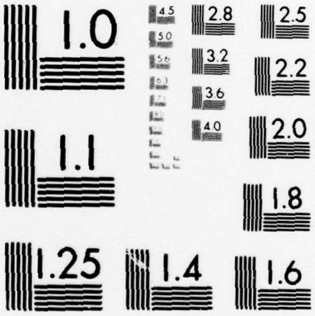
AD-A062 774 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
AN APPROACH TO THE PROGRAMMING OF BIASED REGRESSION ALGORITHMS.(U)
NOV 78 R F GUNST AFOSR-75-2871
UNCLASSIFIED AFOSR-TR-78-1548 NL

| OF |
AD
A062774

END
DATE
FILMED
3-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A062774

DDC FILE COPY

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19 REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 AFOSR TR- 7 ~~7~~8-1548 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER LEVEL II
4. TITLE (and Subtitle) 6 AN APPROACH TO THE PROGRAMMING OF BIASED REGRESSION ALGORITHMS. 2	5. TYPE OF REPORT & PERIOD COVERED 9 Interim rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) 10 Richard F./Gunst	8. CONTRACT OR GRANT NUMBER(s) 15 AFOSR-75-2871
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Department of Statistics Dallas, Texas 75275	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 16 17 A5 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	12. REPORT DATE 11 Novem~~ber 1~~978
	13. NUMBER OF PAGES 10 12 11 p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear Regression
Biased Estimation
Computer Software

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Due to the near nonexistence of computer algorithms for calculating estimators and ancillary statistics that are needed for biased regression methodologies, many users of these methodologies are forced to write their own programs. Brute-force coding of such programs can result in a great waste of computer core and computing time, as well as inefficient and inaccurate computing techniques. This article proposes some guides to more efficient programming by taking advantage of mathematical similarities among several of the more popular biased regression estimators.

404 198

JDB

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AFOSR-TR- 78-1548

AN APPROACH TO THE PROGRAMMING OF BIASED REGRESSION ALGORITHMS

Richard F. Gunst

Southern Methodist University
Dallas, Texas 75275

ABSTRACT

Due to the near nonexistence of computer algorithms for calculating estimators and ancillary statistics that are needed for biased regression methodologies, many users of these methodologies are forced to write their own programs. Brute-force coding of such programs can result in a great waste of computer core and computing time, as well as inefficient and inaccurate computing techniques. This article proposes some guides to more efficient programming by taking advantage of mathematical similarities among several of the more popular biased regression estimators.

1. INTRODUCTION

Regression data analysts currently face a serious computing problem in their efforts to utilize biased regression techniques. On the one hand, there is a vast amount of evidence in scientific publications that biased regression procedures are preferable to ordinary least squares estimation when the predictor variables are multicollinear (e.g., Dempster, Schatzoff, and Wermuth (1977) and

Gunst and Mason (1977b)). Ridge Regression (Hoerl and Kennard (1970)), Principal Component Regression (Massy (1965), Marquardt (1970)), Latent Root Regression (Hawkins (1973), Webster, Gunst, and Mason (1974)), and Shrunken Estimators (James and Stein (1961), Mayer and Willke (1973)) encompass a wide variety of popular biased regression methodologies that have been proposed as alternates to unbiased least squares estimation.

Countering the avowed need for biased regression techniques, on the other hand, is a dearth of computer programs in the standard program libraries (BMDP (Dixon, 1975), SPSS (Nie, et al., 1975), etc.) that the data analyst can access to perform the required calculations. Many users of biased regression techniques, given the time lag between the advent of new biased regression procedures and the introduction of appropriate computer software, are forced to code their own algorithms. Most of these users are not primarily computer programming experts but acquire sufficient knowledge of a programming language such as FORTRAN to be able to write software needed in their research. It is to these users that this article is addressed.

The general theme of this article is a discussion of similarities inherent in the biased estimators listed above and some of the more useful diagnostic measures as well. Biased regression methodologies employ estimators which, although appearing quite different, can be expressed as functions of common variables. Some of these estimators are so similar when reexpressed in terms of these common variables that several authors have grouped them into "families" (e.g., Hocking, Speed, and Lynn (1976), Gunst and Mason (1977b)). By taking advantage of the mathematical similarities of the estimators, core storage requirements and computing time can be lessened.

2. INPUT / DIAGNOSTICS

The basic input to a regression program is an (n × 1) raw response vector, $\underline{Y}^*$, and an (n × p) raw data matrix of predictor variables, $X^* = [X^*_{ij}]$. Large core requirements can be necessitated if $\underline{Y}^*$ and X^* are to be stored and retained during all program calculations. For virtually all the computations except the calculation of residuals, however, only summary statistics and pairwise correlations of the (p+1) input variables are needed. Thus only these statistics need be stored by the program. The elements of $\underline{Y}^*$ and X^* can be stored on peripheral mass storage devices and only called for during initial calculations and the computation of residuals; when not needed, the arrays can be returned to the peripheral storage units.

It is well-documented that for most regression computations some form of standardization is desirable (e.g., Marquardt and Snee (1975)). Let $\underline{Y}$ and X denote the "unit length" standardization of $\underline{Y}^*$ and X^*:

$$Y_i = (Y^*_i - \bar{Y}^*)/d_y \qquad X_{ij} = (X^*_{ij} - \bar{X}^*_j)/d_j$$

$$\bar{Y}^* = n^{-1} \sum_{i=1}^{n} Y^*_i \qquad \bar{X}^*_j = n^{-1} \sum_{i=1}^{n} X^*_{ij}$$

$$d_y = \{ \sum_{i=1}^{n} (Y^*_i - \bar{Y}^*)^2 \}^{1/2} \qquad d_j = \{ \sum_{i=1}^{n} (X^*_{ij} - \bar{X}^*_j)^2 \}^{1/2}.$$

Arrays containing the means, $\bar{Y}^*$ and $\bar{X}^*_j$, root sums of squared deviations, d_y and d_j, correlations between the response and predictor variables, elements of $X'\underline{Y}$, and correlations between pairs of predictor variables, elements of $X'X$, then contain the information needed for the calculation of biased regression estimators. These arrays also contain valuable diagnostic information regarding associations among the predictor variables.

Routinely, the means and standard deviations of the input variables and the arrays $X'\underline{Y}$ and $X'X$ should be output for regression data. The means and standard deviations yield summary information about the location and dispersion of the input variables which can aid in assessing whether the data collected is representative of the process or phenomenon under study. Pairwise correlations indicate the strength of linear associations between two variables. In particular, large pairwise correlations among the predictor variables alert the user to the possibility of strong multicollinearities which might have an adverse effect on least squares estimation and variable selection techniques (for a survey of the problems associated with multicollinearities, see Mason, Gunst, and Webster (1975)).

Latent roots and vectors of X'X provide additional information on multicollinearities, particularly multicollinearities involving more than two predictor variables (and, as we shall see in the next section, form one basis for the expression of biased estimators as a family). Define the latent roots, $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_p$, and the corresponding latent vectors, $\underset{\sim}{V}_1, \underset{\sim}{V}_2, \ldots, \underset{\sim}{V}_p$, of X'X by

$$(X'X - \ell_j I)\underset{\sim}{V}_j = \underset{\sim}{0} \qquad j = 1, 2, \ldots, p \ .$$

Latent vectors corresponding to latent roots that are near zero identify multicollinearities among the predictor variables. Specifically, large elements of these latent vectors indicate which variables are involved in multicollinearities and the nature of the individual multicollinearities (for a detailed illustration of the use of latent roots and vectors in the detection of multicollinearities see Gunst and Mason (1977a)).

An additional diagnostic measure that is useful in assessing multicollinearities is the variance inflation factor (VIF) of each predictor variable (Marquardt (1970), Marquardt and Snee (1975)). The VIF of the jth predictor variable is the jth diagonal element of $(X'X)^{-1}$. If X is an orthogonal matrix all the VIF equal 1.0 since $X'X = (X'X)^{-1} = I$, the $(p \times p)$ identity matrix. The more multicollinear the predictor variables, the larger are the VIF for the variables involved in the multicollinearities. Values of the VIF larger than 10, or even as large as 6, indicate strong multicollinearities and potential difficulties with least squares estimation.

Rather than computing $(X'X)^{-1}$ from a separate algorithm in order to obtain the VIF, the latent roots and vectors of X'X can be used instead. From the relationship

$$X'X = VLV' = \sum_{r=1}^{p} \ell_r \underline{V}_r \underline{V}_r' \quad , \tag{2.1}$$

it follows immediately that

$$(X'X)^{-1} = VL^{-1}V' = \sum_{r=1}^{p} \ell_r^{-1} \underline{V}_r \underline{V}_r' \quad , \tag{2.2}$$

where $V = [\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_p]$ and $L = \mathrm{diag}(\ell_1, \ell_2, \ldots, \ell_p)$. Thus if $C = (X'X)^{-1}$, the jth VIF is given by

$$C_{jj} = \sum_{r=1}^{p} \ell_r^{-1} V_{jr}^2 \quad . \tag{2.3}$$

By taking advantage of the mathematical property (2.1), there is no need to compute nor store $(X'X)^{-1}$ once the latent roots and vectors of X'X are obtained.

Other informative summary and diagnostic information such as the minimum and maximum of each input variable, two variable plots, or measures of how influential each data point is on the estimation of the regression coefficients (e.g. Cook (1977)) could also be computed or available as optional output. Any or all of these diagnostic measures could be indispensable for proper analysis and interpretation of a regression data set. All should be available to the user.

3. ESTIMATORS

The five estimators mentioned in the Introduction are defined mathematically in the following equations, all of which employ standardized input variables. Least squares (LS) estimators are given by

$$\hat{\underline{\beta}}_{LS} = (X'X)^{-1}X'\underline{Y}^* = d_y(X'X)^{-1}X'\underline{Y} \quad . \tag{3.1}$$

For some k > 0, (simple) ridge regression (RR) estimators can be written as

$$\hat{\underline{\beta}}_{RR} = d_y(X'X + kI)^{-1}X'\underline{Y} \quad . \qquad (3.2)$$

A principal component (PC) estimator which deletes the first s components (obvious alterations can be made if subsets other than the first s are to be deleted) can be obtained as

$$\hat{\underline{\beta}}_{PC} = d_y(X'X)^{+}X'\underline{Y} \quad , \qquad (3.3)$$

where $(X'X)^+ = VL^+V'$ and $L^+ = \text{diag}(0, 0, \ldots, 0, \ell_{s+1}^{-1}, \ell_{s+2}^{-1}, \ldots, \ell_p^{-1})$. Shrunken estimators (SE) can be calculated by

$$\hat{\underline{\beta}}_{SE} = g\hat{\underline{\beta}}_{LS} = gd_y(X'X)^{-1}X'\underline{Y} \quad , \qquad (3.4)$$

where $0 \leq g \leq 1$. Finally, latent root estimators (LR) are functions of the latent roots, $\lambda_o \leq \lambda_1 \leq \cdots \leq \lambda_p$, and the corresponding latent vectors, $\underline{\gamma}_o, \underline{\gamma}_1, \ldots, \underline{\gamma}_p$, of the (p+1) by (p+1) matrix A'A, where $A = [\underline{Y}:X]$. (This matrix is already available from the initial arrays since

$$A'A = \begin{bmatrix} 1 & \underline{Y}'X \\ X'\underline{Y} & X'X \end{bmatrix}$$

and the same algorithm used to calculate the latent roots and vectors of X'X can be used to calculate those of A'A). For ease of notation let $\underline{\gamma}_j' = (\gamma_{oj}:\underline{\delta}_j')$ where $\underline{\delta}_j' = (\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{pj})$. Then the latent root estimator can be written as

$$\hat{\underline{\beta}}_{LR} = d_y \sum_r f_r \underline{\delta}_r \quad , \qquad (3.5)$$

where $f_r = -\gamma_{or}\lambda_r^{-1}/(\sum_q \gamma_{oq}^2 \lambda_q^{-1})$ and the summations are taken over all subscripts for which γ_{oj} and λ_j are not simultaneously close to zero.

Equations (3.1) to (3.5) appear to indicate that several matrix inversions and large storage requirements are needed to calculate all the biased estimators listed. Actually, apart from the initial arrays mentioned in Section 2, only the latent roots and vectors of X'X and A'A need be computed and stored. All five estimators can be expressed in the general form

$$\hat{\underline{\beta}} = d_Y \sum_r h_r \underline{m}_r \quad , \tag{3.6}$$

where the h_r are appropriately defined univariate variables and the $\underline{m}_r$ are latent vectors of either X'X or A'A. Specifically, h_r and $\underline{m}_r$ are defined as follows for the five estimators:

$$\begin{array}{llll}
\text{LS:} & \underline{m}_r = \underline{v}_r \ , & h_r = \ell_r^{-1}\underline{V}_r'X'\underline{Y} & r = 1,2,\ldots,p \\[1ex]
\text{RR:} & \underline{m}_r = \underline{v}_r \ , & h_r = (\ell_r + k)^{-1}\underline{V}_r'X'\underline{Y} & r = 1,2,\ldots,p \\[1ex]
\text{PC:} & \underline{m}_r = \underline{v}_r \ , & h_r = \begin{cases} 0 \\ \ell_r^{-1}\underline{V}_r'X'\underline{Y} \end{cases} & \begin{array}{l} r = 1,2,\ldots,s \\ r = s+1,\ldots,p \end{array} \\[1ex]
\text{SE:} & \underline{m}_r = \underline{v}_r \ , & h_r = g\ell_r^{-1}\underline{V}_r'X'\underline{Y} & r = 1,2,\ldots,p \\[1ex]
\text{LR:} & \underline{m}_r = \underline{\delta}_r \ , & h_r = \begin{cases} 0 \\ f_r \end{cases} & \begin{array}{l} \gamma_{or} \approx 0 \text{ and } \lambda_r \approx 0 \\ \text{otherwise .} \end{array}
\end{array} \tag{3.7}$$

Not only are large core storage requirements reduced by using (3.6) and (3.7) since $(X'X)^{-1}$, $(X'X + kI)^{-1}$, and $(X'X)^+$ do not need to be retained, but computing time is shortened in at least two ways. First, $\underline{V}_j'X'\underline{Y}$ appears in several of the h_r in (3.7) but each of these p variables need only be computed once. Secondly, if one wishes to examine several choices of k for RR or several selections of s for PC, for example, repeated calculation of $(X'X + kI)^{-1}$ and $(X'X)^+$ and then $\hat{\underline{\beta}}_{RR}$ and $\hat{\underline{\beta}}_{PC}$ through (3.2) and (3.3) need not be accomplished. It is computationally quite simple and relatively fast to alter k and s in (3.7) and calculate the estimators using (3.6).

4. CONCLUDING REMARKS

Other useful statistics such as variable selection measures can be expressed uniformly just as the estimators in the previous section. One should seek such expressions when writing statistical software in order to take advantage of reduced storage and computing time capabilities. Not only will reductions in storage and computing time result in monetary savings, but the data analyst will find that the computer programs so written will also be able to process much larger data sets than if the suggestions made in this paper were not followed. Several hundred observations on a moderate amount of predictor variables can be a prohibitively large number if $\underline{Y}^*$, X^*, $(X'X)^{-1}$, $(X'X + kI)^{-1}$, etc. must be stored for each computing run.

ACKNOWLEDGMENTS

This research was sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-75-2871. The United States Government is authorized to distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

BIBLIOGRAPHY

Cook, R.D. (1977). "Detection of Influential Observations in Linear Regression," Technometrics, 19, 15-18.

Dempster, A.P., Schatzoff, M., and Wermuth, N. (1977). "A Simulation Study of Alternatives to Ordinary Least Squares," Journal of the American Statistical Association, 72, 77-90.

Dixon, W.J., ed. (1975). BMDP-Biomedical Computer Programs. Berkeley: University of California Press.

Gunst, R.F. and Mason, R.L. (1977a). "Advantages of Examining Multicollinearities in Regression Analysis," Biometrics, 33, 249-60.

Gunst, R.F. and Mason, R.L. (1977b). "Biased Estimation in Regression: An Evaluation Using Mean Squared Error," Journal of the American Statistical Association, 72, 616-628.

Hawkins, D.M. (1973). "On the Investigation of Alternative Regressions by Principal Component Analysis, "Applied Statistics, 22, 275-86.

Hocking, R.R., Speed, F.M., and Lynn, M.T. (1976). "A Class of Biased Estimators in Linear Regression," *Technometrics*, *18*, 425-38.

Hoerl, A.E. and Kennard, R.W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, *12*, 55-67.

James, W. and Stein, C. (1961). "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 361-79.

Marquardt, D.W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, *12*, 591-612.

Marquardt, D.W. and Snee, R.D. (1975). "Ridge Regression in Practice," *The American Statistician*, *29*, 3-20.

Mason, R.L., Gunst, R.F., and Webster, J.T. (1975). "Regression Analysis and Problems of Multicollinearity," *Communications in Statistics*, *4*, 277-92.

Massy, W.F. (1965). "Principal Component Regression in Exploratory Statistical Research," *Journal of the American Statistical Association*, *60*, 234-56.

Mayer, L.S. and Willke, T.A. (1973). "On Biased Estimation in Linear Models," *Technometrics*, *15*, 497-508.

Nie, N., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. (1975). *SPSS-Statistical Package for the Social Sciences*. New York: McGraw-Hill Book Co.

Webster, J.T., Gunst, R.F., and Mason, R.L. (1974). "Latent Root Regression Analysis," *Technometrics*, *16*, 513-22.

KEY WORDS

Linear Regression
Biased Estimation
Computer Software